The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1295–1298
DOI: 10.1214/09-AOAS312F
Main article DOI: 10.1214/09-AOAS312

DISCUSSION OF: BROWNIAN DISTANCE COVARIANCE

By Bruno Rémillard

HEC Montréal

In Székely, Rizzo and Bakirov (2007), the notions of distance covariance and distance correlation between two random vectors were introduced. It was shown that the distance covariance is zero if and only if the two vectors were independent. An empirical version was also defined and its limiting distribution was investigated, under the null hypothesis of independence; furthermore, the underlying test based on the empirical version of the distance covariance is consistent in the sense that under the hypothesis of dependence, its power tends to one as the sample size tends to infinity.

In the present paper the authors continue the study of the properties of the distance covariance and they show that it can be defined in terms of covariances of multivariate Brownian processes. They also generalized that idea to other stochastic processes, namely, multivariate fractional Brownian motions. Defining dependence measures through other stochastic processes is quite interesting, but except for the few cases stated in the paper, it is still to be proven useful. I encourage the authors to continue to explore that interesting idea. Here are some questions I would like to be answered: (i) Can other dependence measures be written in that form, for example, Kendall's tau? (ii) What are the conditions on the underlying processes so that the value of the covariance is zero if and only if the two random vectors are independent? (iii) Can you prove a central limit theorem for the empirical version and what are the conditions on the underlying stochastic processes for the existence of the limiting distribution?

In what follows I will suggest some other extensions and applications of the notion of covariance distance and distance correlation. More precisely, I will describe extensions using rank-based methods and suggest two applications in a multivariate context, that is, when more than two random vectors are involved.

1. Rank-based methods.

In my opinion, there are two weaknesses of the distance covariance: The moment assumption on the random vectors and the

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Applied Statistics*, 2009, Vol. 3, No. 4, 1295–1298. This reprint differs from the original in pagination and typographic detail.

fact that the dependence measure depends on the marginal distributions. That problem can be dealt with easily when the margins are continuous by using the associated uniform variables defined through the well-known mapping

$$X^{(j)} \mapsto U^{(j)} = F_{X^{(j)}}(X^{(j)}), \qquad j = 1, \ldots, p,$$

$$Y^{(k)} \mapsto V^{(k)} = F_{Y^{(k)}}(Y^{(k)}), \qquad k = 1, \ldots, q.$$

Then, the distance covariance between $U = (U^{(1)}, \ldots, U^{(p)})$ and $V = (V^{(1)}, \ldots, V^{(q)})$ only depends on the underlying copula of (U, V) and X and Y are independent if and only if U and V are independent. Its empirical counterpart is simply computed by replacing the observations by their normalized ranks, that is, replacing (X_i, Y_i) by $(R_{X,i}/n, R_{Y,i}/n)$, where $R_{X,ij}$ is the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$, $j = 1, \ldots, p$. It is relatively easy to prove that the limiting distribution of $n\mathcal{V}_n^2(U, V)$ will converge to $\|\xi\|^2$, where the covariance of ξ is $R_{U,V}$, as has been defined in Theorem 5.

On the subject of rank-based methods, I disagree with the authors when they say that these methods are effective only for testing linear or monotone types of dependence. Because independence can also be characterized by copulas, and the latter can be efficiently estimated with ranks, their statement is totally inadequate. See, for example, Genest and Rémillard (2004) for tests of nonserial and serial dependance based on ranks. Furthermore, in their Example 2, the authors suggest that the test based on the distance covariance is more powerful that its rank-based analog. Looking at Figure 2, this is the case only when the sample size n is quite small ($n \le 15$). I would be more convinced by a simulation with different dependence models and sample sizes of the order 50 or 100, at the very least.

2. Measuring dependence between several random vectors. As a competitor to the distance covariance for tests of independence, it is worth mentioning the Cramér–von Mises statistic nB_n, where

$$B_n = \int_{\mathbb{R}^{p+q}} \{F_{X,Y}^n(x, y) - F_X^n(x)F_Y^n(y)\}^2 \, dF_{X,Y}^n(x, y)$$

is the empirical counterpart of

$$B = \int_{\mathbb{R}^{p+q}} \{F_{X,Y}(x, y) - F_X(x)F_Y(y)\}^2 \, dF_{X,Y}(x, y).$$

The latter dependence measure also characterizes independence in the sense that $B = 0$ only when X and Y are independent.

The limiting distribution of $n^{1/2}\{F_{X,Y}^n(x, y) - F_X^n(x)F_Y^n(y)\}$ used to construct B_n was studied in Beran, Bilodeau and Lafaye de Micheaux (2007).

In fact, the authors proposed testing independence between d random vectors $Z_1, \dots, Z_d$, using statistics based on $\mathbb{F}_n = n^{1/2}\{H_n(z_1, \dots, z_d) - F_{n,1}(z_1) \cdots F_{n,d}(z_d)\}$, where H_n is the empirical joint distribution function of $(Z_1, \dots, Z_d)$, and $F_{n,j}$ is the empirical joint distribution of Z_j, $j \in \{1, \dots, d\}$, calculated from a sample $(Z_{11}, \dots, Z_{1d}), \dots, (Z_{n1}, \dots, Z_{nd})$. Extending the results of Ghoudi, Kulperger and Rémillard (2001) from random variables to random vectors, Beran, Bilodeau and Lafaye de Micheaux (2007) considered tests of nonserial and serial dependence based on Möbius decomposition of $\mathbb{F}_n$, yielding asymptotically independent empirical processes $\mathbb{F}_{n,A}$ (depending only on the indices in A), for any subset A of $\{1, \dots, d\}$ containing at least two elements. These $2^d - d - 1$ processes can be combined to define powerful tests of independence [Genest, Quessy and Rémillard (2007)].

Because the limiting distribution under the null hypothesis depends on the unknown distribution function $F_1, \dots, F_d$, Beran, Bilodeau and Lafaye de Micheaux (2007) showed that bootstrap methods worked for estimating the P-value of underlying test statistics.

Further, note that Bilodeau and Lafaye de Micheaux (2005) defined tests on independence between random vectors based on characteristic functions, when the marginal distributions were assumed to be Gaussian. They considered both serial and nonserial cases. The Cramér–von Mises type statistics they used are quite similar to the statistic $n\mathcal{V}_n^2$, when restricted to two random vectors. Therefore, it would be worth considering distance covariance measures for measuring independence between several random vectors. In order to get nice covariance structures, Möbius transformations of the empirical characteristic functions should be used. More precisely, for any $A \subset \{1, \dots, d\}$, one could define distance covariance measures $\mathcal{V}_{n,A} = \|\xi_{n,A}\|^2$, where

$$\xi_{n,A}(t_1, \dots, t_d) = n^{-1/2} \sum_{j=1}^{n} \prod_{k \in A} \{e^{i\langle t_k, Z_{jk}\rangle} - f_{X_k}^n(t_k)\}.$$

3. Measuring dependence for multivariate time series. The distance covariance measures should also be defined in a time series context to measure serial dependence. For example, if $(Z_i)_{i \ge 1}$ is a stationary multivariate time series, one can easily define the "distance autocovariance" by

$$\mathcal{V}^2(l) = \mathcal{V}^2(Z_j, Z_{j+l}), \qquad l \ge 1.$$

It is easy to show that under the white noise hypothesis and the assumption that $|Z_1|_p$ has finite expectation,

$$n\mathcal{V}_n^2(l) \xrightarrow{D} \|\xi_l\|^2,$$

where $\xi_1, \dots, \xi_m$ are independent copies of ξ, as defined in Theorem 5. Again, Möbius transformations should be used to test independence between

$(Z_1, \ldots, Z_m)$. Therefore, there are still many interesting avenues to explore, especially for time series applications. For example, rank-based methods could also be used. See, for example, Genest and Rémillard (2004).

4. Using residuals and pseudo-observations. Finally, one could ask what happens when observations are replaced by residuals (or pseudo-observations like normalized ranks)? For example, one would like to test independence of the error terms in several linear models, using the residuals. Based on the results in Ghoudi, Kulperger and Rémillard (2001), the limiting distribution of $n\mathcal{V}_n^2$ should remain the same, under weak assumptions. That should also be true for the multidimensional extensions of the distance covariance. However, replacing the unobservable innovations by residuals in multivariate time series models leads to completely different limiting processes. For example, using residuals of a simple AR(1) model of the form $Z_t = \mu + \phi(Z_{t-1} - \mu) + \varepsilon_t$, one can show that $n\mathcal{V}_n^2(l)$ converges in law to $\|\xi_l - \gamma_l\|^2$, where

$$\gamma_l(t,s) = sf(s)f'(t)\Phi\phi^{l-1},$$

where f is the characteristic function of ε_t, and ϕ_n is an estimator of ϕ so that $n^{1/2}(\phi_n - \phi)$ converges in law to Φ.

Fortunately, using an analog of the transform Ψ defined in Genest, Ghoudi and Rémillard [(2007), page 1373], it might be possible to obtain limiting distributions not depending on the estimated parameters.

REFERENCES

Beran, R., Bilodeau, M. and Lafaye de Micheaux, P. (2007). Nonparametric tests of independence between random vectors. *J. Multivariate Anal.* **98** 1805–1824. MR2392434

Bilodeau, M. and Lafaye de Micheaux, P. (2005). A multivariate empirical characteristic function test of independence with normal marginals. *J. Multivariate Anal.* **95** 345–369. MR2170401

Genest, C., Ghoudi, K. and Rémillard, B. (2007). Rank-based extensions of the Brock Dechert Scheinkman test for serial dependence. *J. Amer. Statist. Assoc.* **102** 1363–1376. MR2372539

Genest, C., Quessy, J.-F. and Rémillard, B. (2007). Asymptotic local efficiency of Cramér–von Mises tests for multivariate independence. *Ann. Statist.* **35** 166–191. MR2332273

Genest, C. and Rémillard, B. (2004). Tests of independence or randomness based on the empirical copula process. *Test* **13** 335–370. MR2154005

Ghoudi, K., Kulperger, R. J. and Rémillard, B. (2001). A nonparametric test of serial independence for time series and residuals. *J. Multivariate Anal.* **79** 191–218. MR1868288

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

GERAD and Service de l'enseignement
des méthodes quantitatives de gestion
HEC Montréal
Montréal
Canada H3T 2A7